

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : **11-015604**

(43)Date of publication of application : **22.01.1999**

(51)Int.Cl.

**G06F 3/06**

**G06F 12/16**

(21)Application number : **09-168898**

(71)Applicant : **HITACHI LTD**

(22)Date of filing : **25.06.1997**

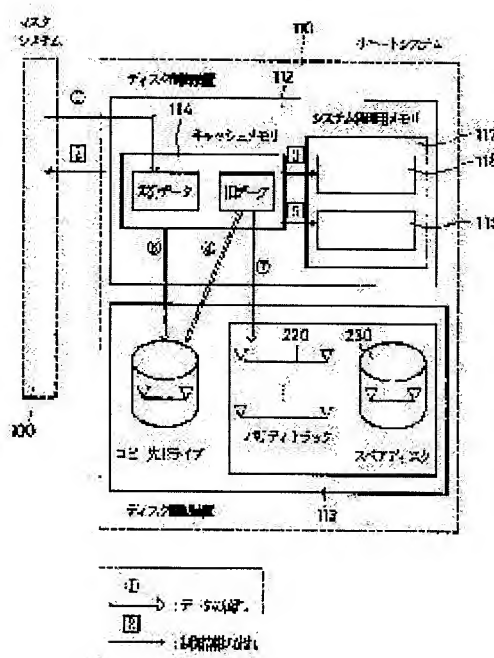
(72)Inventor : **KIMURA YUKIHISA  
NAGASAWA MITSUO  
NAKAMURA KATSUNORI**

## (54) DATA MULTIPLEX METHOD

### (57)Abstract:

**PROBLEM TO BE SOLVED:** To improve the fault resistance of restoration copying for restoring multiplex in a data multiplex system and to maintain the matching of copy data.

**SOLUTION:** In restoration copying for maintaining data multiplex between a master system 100 and a remote system 110, former data on a disk drive unit 113, which corresponds to differential data, is saved to a cache memory 114 or a parity track 220 or a spare disk 230 before overwriting and differential data are stored in the target storage area of the disk drive unit 113 on the side of the remote system 110 receiving differential data from the master system 100. When restoration copying is failed owing to the fault of a master system 100-side, former saved data are returned to the disk drive unit 113 and the matching of copy data is maintained.



## LEGAL STATUS

[Date of request for examination] 16.04.2001

[Date of sending the examiner's decision of rejection] 27.07.2004

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-15604

(43) 公開日 平成11年(1999) 1 月22日

(51) Int.Cl.<sup>6</sup>

G 0 6 F 3/06  
12/16

識別記号

3 0 4  
3 1 0

F I

G 0 6 F 3/06  
12/16

3 0 4 E  
3 1 0 J

審査請求 未請求 請求項の数 3 O L (全 20 頁)

(21) 出願番号 特願平9-168898

(22) 出願日 平成9年(1997) 6 月25日

(71) 出願人 000005108

株式会社日立製作所  
東京都千代田区神田駿河台四丁目6番地

(72) 発明者 木村 恭久

神奈川県小田原市国府津2880番地 株式会  
社日立製作所ストレージシステム事業部内

(72) 発明者 長澤 光男

神奈川県小田原市国府津2880番地 株式会  
社日立製作所ストレージシステム事業部内

(72) 発明者 中村 勝憲

神奈川県小田原市国府津2880番地 株式会  
社日立製作所ストレージシステム事業部内

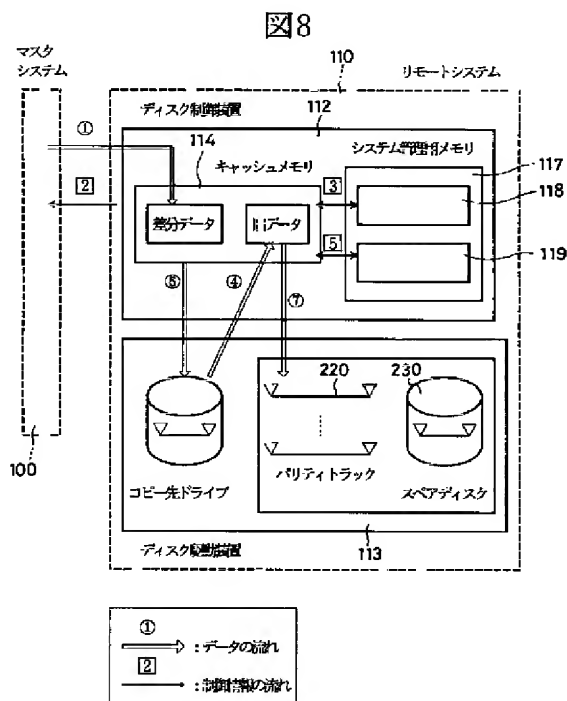
(74) 代理人 弁理士 筒井 大和

(54) 【発明の名称】 データ多重化方法

(57) 【要約】

【課題】 データ多重化システムにおける多重化回復のための回復コピーの障害耐性を向上させ、コピーデータの整合性を維持する。

【解決手段】 マスタシステム100とリモートシステム110との間におけるデータ多重化を維持するための回復コピーにおいて、マスタシステム100から差分データを受け取るリモートシステム110の側では、上書きする前に、当該差分データに対応したディスク駆動装置113上の旧データを、キャッシュメモリ114やパリティトラック220あるいはスペアディスク230に退避させた後、当該差分データをディスク駆動装置113の目的記憶領域に格納する。マスタシステム100側の障害等で回復コピーが失敗した場合には、退避してあった旧データをディスク駆動装置113に戻して、コピーデータの整合性を維持する。



**【特許請求の範囲】**

【請求項1】 第1のシステムに設けられた第1の記憶装置に格納されたデータを、少なくとも一つの第2のシステムに設けられた第2の記憶装置に複写することで前記データの多重化を行うデータ多重化方法であって、任意の要因にて前記第1のシステム側に蓄積された未複写の前記データを前記第2のシステム側に転送する回復複写処理の実行に際して、前記第2のシステム側では、前記第1のシステムから到来する前記データに対応した旧データを任意の方法にて保存しておき、前記回復複写処理が失敗した時には、保存されている前記旧データを用いて前記第1のシステムと前記第2のシステムとの間における前記データの整合性を維持することを特徴とするデータ多重化方法。

【請求項2】 請求項1記載のデータ多重化方法において、

前記第2のシステムでは、

前記第1のシステムから到来する前記データに対応した前記旧データを前記第2の記憶装置における現在の記憶領域から他の任意の記憶領域または他の任意の記憶媒体に退避させることで前記旧データの保存を行う第1の操作、

前記第1のシステムから到来する前記データを、当該データに対応する前記旧データの格納領域とは異なる格納領域または記憶媒体に一時的に保持することで前記旧データの保存を行う第2の操作、

前記回復複写処理の完了までの許容最大時間が任意に設定され、前記許容最大時間を超過しても前記第1のシステム側から前記回復複写処理の完了報告がない場合には、前記回復複写処理が失敗したものと見なし前記旧データによる前記整合性の維持を行う第3の操作、の少なくとも一つの操作を行うことを特徴とするデータ多重化方法。

【請求項3】 請求項1または2記載のデータ多重化方法において、

前記第1のシステムでは、任意の前記要因によって、当該第1のシステム内で発生した前記データを前記第2のシステムに複写できずに蓄積するとき、更新が発生した順序が弁別可能に前記データを蓄積し、前記回復複写処理の実行時には、最も過去に発生した前記データから順に時系列に前記第2のシステムに転送する第4の操作を行うことを特徴とするデータ多重化方法。

**【発明の詳細な説明】****【0001】**

【発明の属する技術分野】本発明は、データ多重化技術に関し、特に、例えば遠隔地に設置された複数のシステム間にてデータを複写することでデータ多重化を実現する場合におけるデータの整合性の維持管理等に適用して有効な技術に関する。

**【0002】**

【従来の技術】情報処理システムは、社会活動の広汎な領域に利用されており、その故障は重大な社会的混乱を招く要因となり得る。このため、例えば、取り扱うデータの喪失が許されない金融等の分野を初めとして、以前より、取扱うデータを失うことのないよう、運用システム配下において、常時バックアップデータを採取保存することは基本的に行われてきた。

【0003】ところが、最近の天災や事故の教訓から、通常運用しているマスタシステムが復旧不可能な場合に陥り、大量のバックアップデータやログからのデータ修復に要する手間や時間、更には情報を失うことなどの問題点を考慮し、遠隔地に設置されているリモートシステムにバックアップコピーデータを貯えるための方法が、例えば特平8-509565号公報に開示された「遠隔データのミラー化」等の技術として新たに提唱された。

【0004】その方法は、マスタシステムのデータをそのままリモートシステムに反映することにより「ミラー状態」を維持するものであり、マスタシステムでの運用が不可能となった場合に、リモートシステムに運用を移行し、システム運用再開を一段と容易にしようとするものである。

【0005】また、リモートシステムへのバックアップコピー方法としては、マスタシステムとリモートシステム間のデータの更新契機から、大きく「同期型」と「非同期型」の2種類に分けることができる。

【0006】前者は、マスタシステムのホストより発生した各更新I/O命令に対し、まずマスタサイドの記憶装置に書き込みを行い、続いてリモートシステムの記憶装置に向けて書き込みを実施し、リモートシステムより書き込み終了通知を受領したところで、マスタサイドの記憶装置はマスタシステムのホストに対して、最終的な書き込み終了報告を行い、常にマスタサイドとリモートサイドの更新は同期を保つ技術である。

【0007】それに対し、後者は、マスタシステムのホストより発生した各更新I/Oに対し、マスタサイドの記憶装置に書き込みを終了した時点で、マスタシステムのホストに書き込み終了報告を行い、この更新I/Oに対するリモートシステムへの更新は遅れて、即ち、非同期に実施される技術である。

【0008】通常の「ミラー状態」を維持している条件にあって、仮にマスタシステムが重大災害等で運用不可能に陥った場合、「同期型」においては、データ更新の同期性から「ミラー状態」は維持される。この場合、最終更新に関するデータ分は、タイミングによりリモートシステムに反映されないかもしれないが、整合性に関しては問題ない。一方、「非同期型」においては、リモートシステムには非同期に更新データが反映されるため、場合によっては「ミラー状態」どころか整合性の保証もできなくなる。

【0009】この整合性の維持の方法としては、例えば、特開平6-290125号公報に開示されているように、更新データを待ち行列化して、更新があった順にリモートシステム側のバックアップ記憶媒体上に反映を行う技術が挙げられる。

【0010】

【発明が解決しようとする課題】上述のような従来の技術でのリモートシステムへのバックアップコピー方法は、常にリモートシステムが正常に動作しており、常時バックアップコピーが可能な条件が成立する前提でのみ、リモートバックアップコピーとしての意味をもつものである。

【0011】具体的に例を挙げて説明すると、リモートシステムへのバックアップ運用を実施している最中に、マスタシステムとリモートシステムを結ぶデータ転送路において自動的あるいは、センタ運用者や保守員等による人手介入によって修復可能なレベルの一時的障害（簡単な例を挙げれば、データ転送路としてのケーブルがはずれたケースなど）が発生した場合、その一時的障害が取り除かれるまでの期間は、リモートシステムへのデータの反映が不可能となる（この状態を「サスペンド状態」と呼ぶ）。

【0012】もちろん、この「サスペンド状態」は短期的な場合もあれば、数時間以上を費やす長期的な場合もあり得る。

【0013】一方、マスタシステムにおいては「サスペンド状態」中もデータの更新は、絶え間なく次々と行われる（この「サスペンド状態」中に発生する各更新データを「差分データ」と呼び、総称して「差分データ群」と呼ぶ）。

【0014】その後、「サスペンド状態」が解除され、バックアップコピーが再開可能なノーマルな状態に戻ると、マスタシステムはリモートシステムに向けて、「差分データ」の反映を行いつつ、その間においても逐次入る更新データに対する反映も考慮にいれ、データの「ミラー状態」への回復の為の試行（本試行を「回復コピー」と呼ぶ）を実施する。

【0015】ところが、この「回復コピー」は、バックアップコピーとしての意味を失う程の重大な問題を抱えている。

【0016】一般に、マスタシステム側では次々と入ってくる「差分データ」をマスタシステム側の記憶媒体上に書き込み、記憶媒体領域全体をビットマップ形式で表現し、更新のあった領域に対応するビットを立てることにより、「差分データ」の格納領域を表現した「差分ビットマップ」にて管理している。ところが、複数の「差分データ」：データ1, 2, 3, . . . が存在した場合、「差分ビットマップ」では、更新の行われた順番まではわからない。そのため、「回復コピー」が始まると、「差分ビットマップ」を最初から検索し、差分のあ

る領域をみつけて、そのデータをリモートシステムに反映している。そのため、「回復コピー」の最中にマスタシステムが重大障害（自然災害、人為災害）等によって、データの読み出しが出来ないなどの運用不可状態に陥った場合、リモートシステムでは、運用上情報の整合性を保証する必要がある単位（それは、1つのデータセット単位であったり、複数のお互いに関連のあるデータセットで構成される単位であったりする。その他、物理的あるいは論理的に定義されている各デバイスも1つの単位として考えられる。）での整合性の保証をすることができないため、この単位でのバックアップコピーは運用出来ず、マスタシステムの破壊状態によっては最悪の場合、データ喪失の状態に陥る。

【0017】本発明の目的は、複数のシステム間でのデータ多重化を維持するための差分データの回復複写における障害耐性を向上させることにある。

【0018】本発明の他の目的は、マスタシステムおよびリモートシステムが、ある整合性の維持を要求する情報単位において、特に「サスペンド状態」に陥る瞬間まで「ミラー状態」を維持していた場合の、「回復コピー」におけるリモートシステム側での多重化データの整合性を保証する技術を提供することにある。

【0019】本発明の他の目的は、データ多重化を行うマスタシステムおよびリモートシステムにおいて、マスタシステム側の障害等に際してリモートシステムに運用を切り替える上で運用可能な整合性の保証されたデータを提供することにある。

【0020】

【課題を解決するための手段】本発明は、第1のシステムに設けられた第1の記憶装置に格納されたデータを、少なくとも一つの第2のシステムに設けられた第2の記憶装置に複写することでデータの多重化を行うデータ多重化方法において、任意の要因にて第1のシステム側に蓄積された未複写のデータを第2のシステム側に転送する回復複写処理（「回復コピー」）の実行に際して、第2のシステム側では、第1のシステムから到来するデータに対応した旧データを任意の方法にて保存しておき、回復複写処理が失敗した時には、保存されている旧データを用いて第1のシステムと第2のシステムとの間におけるデータの整合性を維持するものである。

【0021】すなわち、より具体的には、一例として、「同期方式」のデータ多重化にてマスタシステム（第1のシステム）とリモートシステム（第2のシステム）との間におけるデータの「ミラー状態」を維持する場合に、リモートシステム側において、マスタシステム側と「ミラー状態」を維持していた最近、即ち「サスペンド状態」に陥った直前のデータを、「回復コピー」実行の間、別の記憶領域に一時的に退避して保存するものである。

【0022】なお、旧データの保存方法としては、「回

復コピー」実行の間にリモートシステムに到来するバックアップ対象データを、「回復コピー」が完了するまでの間には対応する旧データの上書きせずに、別の記憶領域に一時的にストアし、リモートシステム側の第2の記憶装置上には「サスペンド状態」に陥った直前の整合性を維持した状態の旧データを残しておく方法でもよい。

【0023】また、「回復コピー」中の障害の発生に備えて、マスタシステム側では、リモートシステムが「サスペンド状態」に陥ってから「回復コピー」実行に至るまでの各「差分データ」を、例えば時系列キューを構築して管理することにより発生順序に従って保持する。そして、最も過去にキューに組み入れた差分データから、「回復コピー」を始めることによって、リモートシステム側は、特に差分データの整合性を意識することなしに、「回復コピー」に障害が発生した場合には、当該障害直前の旧データを採用することで、バックアップデータの整合性を保つことができる。

【0024】また、「回復コピー」中におけるマスタシステムおよびマスタシステムとリモートシステムとの間におけるデータ転送経路等の障害の有無をリモートシステム側から的確に判別すべく、「回復コピー」の完了までの許容最大時間を設定する機能を設け、予め、設定された許容最大時間を超過してもマスタシステム側から「回復コピー」の完了報告がない場合には、リモートシステム側では、マスタシステムが重大障害に陥り、運用不可の状態になっていると判断して、保存されていた旧データのレベルで整合性をとる等の整合性復旧のための処理を自動的に開始する構成とすることもできる。

【0025】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照しながら詳細に説明する。

【0026】図1は、本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムの構成の一例を示す概念図である。

【0027】本実施の形態の情報処理システムは、マスタシステム100(M-SYS)と、このマスタシステム100が取扱うデータのバックアップシステムとして機能するリモートシステム110(R-SYS)からなる。

【0028】マスタシステム100は、ホストCPU101、ディスク制御装置102およびディスク駆動装置103の組み合わせからなる情報記憶システム一列を構成しており、同様にリモートシステム110においても、ホストCPU111、ディスク制御装置112およびディスク駆動装置113の組み合わせからなる情報記憶システム一列を構成している。

【0029】ディスク制御装置102とディスク制御装置112の間は、各々に設けられたコネクタ106、コネクタ116を介して、例えば専用回線や公衆回線等の情報通信網等からなるデータ転送路120にて接続され

ている。

【0030】ディスク駆動装置103、ディスク駆動装置113としては、一例として、例えば磁気ディスク、光ディスク、光磁気ディスク、デジタルビデオディスク(DVD)等、任意の記憶媒体を用いた記憶装置を用いることができる。

【0031】ホストCPU101は、ディスク駆動装置103のディスク媒体へのデータ書き込み命令の発行をはじめ、ディスク制御装置102を制御する為の指示を発行し、ディスク制御装置102は、ホストCPU101からの指示を解析し、その結果に基づいて制御を行なう。例えば、ホストCPU101よりデータの書き込み指示があり、そのデータが送られてくると、ディスク制御装置102はその指示に従い、目的のディスク駆動装置103にデータを書込む。更に、ディスク制御装置102は、リモートシステム110の側に向けて、データのバックアップに関するリモートシステム110への制御指示、およびバックアップデータを転送する機能を有する。また、リモートシステム110側のディスク制御装置112は、マスタシステム100からの指示やバックアップデータの受け付けを実施し、書き込みデータをディスク駆動装置113のディスク媒体上に反映する機能を有する。

【0032】また、マスタシステム100のディスク制御装置102およびディスク駆動装置103、そして、リモートシステム110のディスク制御装置112およびディスク駆動装置113は、各々、キャッシュメモリ104、キャッシュメモリ114、更には、ディスク制御装置102、ディスク制御装置112に対する様々な情報や機能(例えばバックアップコピーの為の各種条件やシステム内の構成情報等の設定や指示機能)を提供するサービスプロセッサ105(SVP)、サービスプロセッサ115(SVP)を持つ。

【0033】また、本実施の形態の場合、一例として、ディスク駆動装置103およびディスク駆動装置113は、RAID5のアーキテクチャを有する。

【0034】尚、本実施の形態においては、一例として上述の通り構成されるが、本発明においては、記憶装置としては、ディスク駆動装置やRAID5を採用することに限らず、一般の記憶装置(RAIDアーキテクチャを有さないディスク装置、半導体記憶装置、磁気テープ装置など)や他のRAIDレベル(RAID0, RAID1, ...)構成に対しても適用可能である。また、ホストCPU101, 111、およびディスク制御装置102, 112、更にSVP105, 115における上述の機能については、各々マスタシステム100および、リモートシステム110内に設けられているならば代用可能である。例えば、ディスク制御装置102、112が自らデータの書き込み指示を発行して、ディスク媒体上への反映を行なったり、バックアップコピーに関

する条件設定をSVPを介してではなく、ホストCPU101やディスク制御装置102自身で設定するケースなどが考えられる。

【0035】図2は、本実施の形態のディスク駆動装置103およびディスク駆動装置113の各々におけるRAID5のアーキテクチャの一例を示す概念図である。

【0036】RAID5のアーキテクチャを有するディスク制御装置/駆動装置とは、多数の小型ディスク200, 201, . . . を記憶媒体として用い、大型ディスク装置をエミュレートしたディスク装置であり、何本かの論理トラック210, 211, . . . に対して、データ修復コードであるパリティコードを格納する1本の論理トラック(パリティトラック220)を用意しており、ある1台の小型ディスクが故障してもパリティトラック220のパリティデータを元に故障ディスク内の全データの修復を自発的にを行い、故障したディスクの代替用の小型ディスク(スペアディスク230)にその回復データを反映することにより、オンラインでのシステム運用継続を可能にするなどの特徴を持っている。

【0037】以上に述べた構成の本実施の形態の情報処理システムにおいて、例えば、整合性を維持する単位として、論理デバイスを選んだ場合のバックアップコピーの手順の一例を説明する。図3は、その手順の一例を図式化して例示した概念図であり、図4はそのフローチャートである。

【0038】マスタシステム100において、ホストCPU101、或いはSVP105などより、論理デバイス単位でマスタシステム100とリモートシステム110の間でペア(マスタシステム100側の1論理デバイス(M-DEV)とリモートシステム110側の1論理デバイス(R-DEV)の組であり、R-DEVは、M-DEVに対するデータのコピー先にあたり、M-DEVのデータは、指定されたR-DEV内に格納される。)310, 311, . . . , 312を形成するよう指示が発生すると(ステップ350)、マスタシステム100は(ディスク制御装置102等の記憶制御装置などを介して)、リモートシステム110に向けて、その旨を指示する命令を発行する(ステップ351)。その正常終了報告をリモートシステム110側より受けると(ステップ352)、マスタシステム100はバックアップコピーを開始し、M-DEV#=Miに入った更新データは、そのペアにあたるR-DEV#=Riに反映を行い、「ミラー状態」を維持する(ステップ353)。なお、この「ミラー状態」とは、あくまでも、真のミラー状態である必要はなく、データの多重化が維持されていればよく、物理的に全く同一のトラックに反映しなくてもかまわない。

【0039】ところが、ケーブル等のデータ転送路120が切断されたり、マスタシステム100側のディスク駆動装置103の媒体障害などでペア間のデータコピー

が一時的に不可能となった、即ち、ペアが「サスペンド状態」に陥った場合、ケーブルの交換、ディスク駆動装置103のオンライン交換(データはパリティデータより再生可能)などにより、障害要因が取り除かれ、ペア間のデータコピーが再開可能になるまでの間、ホストCPU101より発行される更新I/Oに対し、ディスク制御装置102は更新のあった論理デバイス番号(L-DEV#)および論理トラック番号(L-TRK#)を後述のような差分ビットマップ107を利用して記憶する。

【0040】図7は、この差分ビットマップ107の構成の一例を示す概念図である。ここでは、論理デバイス番号を行(論理デバイス#=0を1行目、. . . 論理デバイス#=iをi+1行目)、論理トラック番号を列(論理トラック#=0を1列目、. . . 論理トラック#=jをj+1列目)に配列したものとなっている。例えば、マスタシステム100側にて、論理デバイス#=iの論理トラック#=0に差分(トラック)データが発生すると、(i+1)行、1列目のビット107aを立てることにより差分が当該トラックに存在することを表わしている。

【0041】やがて、ペア間のデータコピーが再開可能となった後、ホストCPU101からの指示、或はSVP105からの指示によって、「回復コピー」を始める。

【0042】最初に、この「回復コピー」において、リモートシステム110側では、マスタシステム100側から転送されて来るデータに対応したディスク駆動装置113内の旧データを、上書きされる前に、他の領域に退避させて保存する場合について説明する。図5および図6は、この場合の「回復コピー」の処理の流れの一例を示したフローチャートであり、図5はマスタシステム側の動作を、図6はリモートシステム側の動作を、それぞれ示している。

【0043】本実施の形態においては、旧データの退避エリアとして、リモートシステム110配下のディスク制御装置112に搭載されているキャッシュメモリ114および、ディスク駆動装置113内のパリティトラック220(パリティエリア)或は、スペアディスク230の記憶領域を利用している。特にパリティトラック220を退避エリアとして使用することはリモートシステム110側のディスク駆動装置113が、「回復コピー」を実施したことを契機にRAID5の機能レベルを失うことを意味するが、それよりも、「回復コピー」中にマスタシステム100が運用不可に陥り、更にリモートシステム110側のディスクデータも保証出来ないという技術的課題を解決する方がはるかに大切である。もちろん、パリティトラック220やスペアディスク230の使用可否等の条件をユーザ側にて選択できる機能は用意する。但し、一般的には、「回復コピー」が終了し

て、各ペアが「ミラー状態」を回復した瞬間に至れば、マスタシステム100のパリティデータを考慮に入れば、リモートシステム110側のディスク駆動装置113のRAID5レベルは確保できる。

【0044】さて、マスタシステム100配下のディスク制御装置102は回復コピー指示のイベントを受けると(ステップ401)、まず、SVP105から回復コピー完了許容最大時間Tを入手し(ステップ402)、さらに、リモートシステム110側の記憶制御装置(ディスク制御装置112)のキャッシュメモリ容量、パリティトラック本数およびスベアディスク数などの構成情報をリモートシステム110より入手する(ステップ403)。

【0045】さらに、ユーザが設定した条件(RAID5レベルの維持、スベアディスク230の使用可否、キャッシュメモリ114の最大占有許容量(しきい値)などがあり、リモートシステム110側よりこれらの条件を設定する場合、マスタシステム100は「回復コピー」を開始する前に、予め入手しておく。また、マスタシステム100側よりこれらの情報を設定する場合は、リモートシステム110側に、これらの情報を与えておき、これらの条件を双方のシステムが同一認識している中で「回復コピー」を実施する)を元に、旧データの退避エリアの容量を算出し、同時に「回復コピー」を実施することが可能なペア数の最大値nを決定する(ステップ404)。

【0046】この最大値nの決定の理由は、「回復コピー」処理の効率の面から、確保できる旧データの退避エリアの容量に対応した推奨値を提供する目的で求めるものであり、単に1論理デバイス単位毎に「回復コピー」を実施しても構わない。

【0047】例えば、リモートシステム110側のディスク駆動装置113の構成例として、64論理デバイスを24台の物理ディスク(小型ディスク)でエミュレートし、小型ディスク4台分のパリティトラック(4パリティグループ)および2台のスベアディスク230からなり、キャッシュメモリ114は考慮にない場合、nとして設定できる最大値は $(64/24) \times (4+2)$ の整数部分をとったもの、即ち、 $n=16$ であり、「回復コピー」を実施する上で、旧データの退避エリアがネックとなることはないといえる。また、キャッシュメモリ114を十分に搭載したシステム環境を整えることが可能ならば、キャッシュメモリ114のみを退避エリアとして使用することも可能である。

【0048】次に、マスタシステム100側で「回復コピー」の為の準備が整った段階で、リモートシステム110側に「回復コピー」開始を指示し(ステップ405)、「回復コピー」にあたって使用する情報(回復コピー完了許容最大時間:Tなど)を送り、指示に対する応答を待つ(ステップ406)。

【0049】一方、リモートシステム110側では、「回復コピー」の指示を受けると(ステップ501)、「回復コピー」開始を認識し、監視タイマ(回復コピー完了許容最大時間Tに対する監視)をスタートさせるなどの準備を行い(ステップ502)、「回復コピー中」のフラグをONにし(ステップ503)、マスタシステム100に準備が整ったことを表わす応答を返し(ステップ504)、差分データの到着を待つ(ステップ505)。

【0050】さて、リモートシステム110より「回復コピー」開始指示に対する応答を受けたマスタシステム100側は、差分ビットマップ107(図7)を参照することにより、差分データがあるトラック(差分トラック)を検索し(ステップ407)、見つかると、そのトラックデータをリモートシステム110側に送るべく、書き込み命令を作成・発行し(ステップ408)、リモートシステム110からの書き込み終了報告を待つ(ステップ409)。

【0051】その書き込み命令受信を検知すると、リモートシステム110側は、バックアップドライブ(ディスク駆動装置113)への格納アドレスを求め(ステップ506)、キャッシュメモリ114上に差分データを格納し(ステップ507)、書き込み終了報告を行う(ステップ508)。

【0052】その後、この差分データをディスク駆動装置113の目的のトラックに上書きする前に、当該トラック上の旧データを退避するために、空きエリア(トラック)を探したのち(ステップ509)、一旦トラック上の旧データをキャッシュ上に退避し、旧データが元格納されていたデバイス番号やトラック番号および退避先などの情報を作成して、記憶しておく(ステップ510)。

【0053】尚、退避データが使用可能キャッシュ許容量を越えない限りは(ステップ511)、キャッシュメモリ114のみを使用して退避を続け、そのしきい値を越えた場合に限り、ディスク駆動装置113上の退避エリアに反映する(ステップ512)。

【0054】以上が、1本分の差分トラックに対する「回復コピー」の処理になるが、このあとも引き続き、マスタシステム100側は、差分トラックを検索して(ステップ410)、リモートシステム110側に書き込み命令を送り、あるペアの全差分データを反映し終わると、そのペアiに対する回復到達報告をリモートシステム110側に行い(ステップ411)、「回復コピー」待ちになっていたペアを追加して(ステップ412)、「回復コピー」指示のあった全てのペアに対する差分データをリモートシステム110側に送り続ける。そして、全ての差分データがリモートシステム110側に反映されたところで「回復コピー」完了報告をリモートシステム110に送り(ステップ413)、マスタシ

システム100側の「回復コピー」に対する処理は完了する。

【0055】さて、リモートシステム110側は、「回復コピー」中は、マスタシステム100側からの書き込み命令受信、マスタシステム100側からの「回復コピー」完了報告および、「回復コピー」実行時間タイムオーバのいずれかを検知すると、それに対応した処理へ移行するようになっており（ステップ505、ステップ514、ステップ515で形成されるループ）、「回復コピー」完了報告を受けると、リモートシステム110側は、「回復コピー」完了を認識して“回復コピー中”のフラグをOFFし（ステップ520）、「回復コピー」処理を完了させる。

【0056】一方、マスタシステム100側からの「回復コピー」完了報告がいつまでたってもなく、「回復コピー」実行タイムオーバに達した場合においては、回復到達報告（図5のステップ411）があったベア以外の、整合性を保証できないリモートシステム110側の論理デバイスに対する退避データの有無をチェックし（ステップ516）、退避データがキャッシュメモリ114上に残っているか否かを判別し（ステップ517）、残っていればキャッシュメモリ114より（ステップ518）、そうでない退避データは、パリティトラック220やスペアディスク230からキャッシュメモリ114上に読み込みを行なって、バックアップを行なっている目的のトラックに旧データを反映し（ステップ519）、整合性復旧の為の処理を実施する。

【0057】次に、旧データの退避方法、および手順の一例について詳細に説明を行なう。図8は、その手順の一例を図式化して例示した概念図であり、図9は、そのフローチャートである。

【0058】まず、リモートシステム110は、マスタシステム100から差分データが転送され、そのデータをディスク制御装置112のキャッシュメモリ114内に格納すると、差分データ格納終了報告をマスタシステム100に対し行なう（ステップ751、ステップ752）。その後、差分データによって上書きされる旧データを退避する為に、システム管理用メモリ117上の退避エリア管理ビットマップ118をアクセスし、空きトラックを検索、退避先を決定する（ステップ753）。

【0059】なお、システム管理用メモリ117の割り当て領域には特に制限はなく、リモートシステム110内の任意のメモリ（例えばキャッシュメモリ114など）上に任意に定義することができる。

【0060】図10は、退避エリア管理ビットマップ118の具体例を例示した概念図である。パリティトラック群、あるいは、スペアディスク1台分を行に、およびトラック番号を列に配列し、各退避先トラック1本1本の使用状況をビットマップ802で表現したものである。場合によっては、パリティトラック220、スペア

ディスク230の使用可否フラグ801（ビットマップの1列目に配置）を持つこともある。図10の例では、スペアディスク#0に対する使用可否フラグ803が立っている為、スペアディスク#0のトラックが退避先として許可されており、更に、その論理トラック番号0、nに対するスペアディスク#0上のエリアは“空き”であることを表わす。

【0061】退避エリアを確保すると、差分データによって上書きされる旧データを退避データとしてキャッシュメモリ114上に読み込み（ステップ754）、併せて、整合性回復実施時に必要な退避データに関する諸情報を作成し、1つのノードとして、システム管理用メモリ117の一部に設けられた退避データ情報キュー119にエントリする（ステップ755）。

【0062】図11は、この退避データ情報キュー119の構造と各ノードに関する諸情報の一例を表わす概念図である。この例では、差分データの反映先である論理デバイスの番号毎に、1つのキュー901、902、...、903を設け、1論理トラック分を単位として1つのノードを割り当てている。尚、本実施の形態では一例として退避データ情報キュー119のようなキュー構造を採用しているが、本発明においては、単純なテーブル構造等、任意のデータ構造を採用することも可能である。

【0063】各ノードには、整合性回復を行なう際に必要な情報、具体的には差分データの反映先にあたる物理デバイス番号950やその格納アドレス（物理トラック番号（セクタ番号）951）、旧データの退避先にあたる（スペア）ディスク番号952、および物理トラック番号（セクタ番号）953等の格納アドレス、および旧データのキャッシュメモリ上格納アドレス954などを含んでいる。

【0064】そして、最後に、差分データを記憶媒体（ディスク駆動装置113）上に、退避旧データをパリティトラック220やスペアディスク230上の退避エリアに反映し、1差分データに対する、旧データの退避は完了する（ステップ756、ステップ757）。なお、このステップ756、ステップ757の操作は、非同期に実施されても構わない。

【0065】次に、リモートシステム110側における「回復コピー」実行時の旧データの保存方法の他の例を説明する。この例では、「回復コピー」実行時に旧データの退避は行わず、マスタシステム100から到来する差分データを本来のディスク駆動装置113の書込領域以外の領域に一時的に保持することで、旧データの保存を行う。この例について、図6を参照して、前述の実施の形態との違いを挙げながら説明を行う。

【0066】違いは3つあり、1つ目は、退避エリアを確保すべく、退避エリア管理ビットマップ118を使って、空きエリア（トラック）を探した後（ステップ50

9に対応)、その次のステップ510は行わずに、退避データ情報キュー119に差分データに対する情報を作成しエントリを行う(ステップ511に対応)。本実施の形態においても、キャッシュメモリ占有許容量(しきい値)を越えない限りは、基本的に退避データ(差分データ)をキャッシュメモリ114上に置いておく。

【0067】2つ目は、「回復コピー」完了報告後の処理であり、全差分データを目的のディスク駆動装置113に反映する必要がある。キャッシュメモリ114上になくパリティトラック220やスベアディスク230上に退避してある差分データは、キャッシュメモリ114の空き容量に応じてキャッシュメモリ114上に読み込みを行い、順次、目的のバックアップを行うべきディスク駆動装置113に差分データの反映を行う。

【0068】また、3つ目は、「回復コピー」実行タイムオーバーになったケース、即ち、整合性維持のための処理の部分であり、回復到達報告を受けたペアに対しては、差分データの反映を実施するが、回復到達報告を受けていないペアに対しては差分データは目的のバックアップを行うべきディスク駆動装置113への反映を行わずに破棄する。

【0069】次に、マスタシステム100の側において「サスペンド状態」の時に自システム内で発生した更新データ(差分データ)を時系列に蓄積する場合に用いられる時系列キューの構築方法の一例について説明する。

【0070】図12は、その時系列キュー108の構造を表わしており、1001、1002、...は時系列キュー108の各ノードにあたる。この時系列キュー108は、マスタシステム100内にあるメモリ(例えばディスク制御装置102のキャッシュメモリ104や図示しないシステム管理用メモリなど)に定義・構築し、少なくとも整合性の維持を図る単位(例えば論理デバイスなど)毎に設ける。各ノードは、差分データの発生順にキュー先頭から並んでおり、例えば、論理トラック単位毎に1個のノードを割り当てる。

【0071】時系列キュー108を構成する個々のノード1001、ノード1002、...は、直前に入った差分データに対する、今まで最後尾に接続されていたノードの先頭をさす前エントリノードポインタ1101と、次に発生した差分データに対するノードの先頭をさす次エントリノードポインタ1106とを有する双方向キューの形式をとっている。なお、前エントリノードポインタ1101を省略した一方向キューの形式を採ってもよいことは言うまでもない。

【0072】更に、個々のノード1001、ノード1002、...は、本ノード制御用情報1102と、当該ノードに割り当てられた差分データの論理トラック番号1103と、差分データの格納先を表わす物理媒体上格納アドレス1104と、キャッシュメモリ上アドレス1105とを含んでいる。

【0073】本ノード制御用情報1102には、例えば当該ノードに対する差分データが、論理トラック全体をフォーマットするための更新(フォーマットライト)データなのか、部分的な更新(パーシャルライト)データなのかを表わす更新タイプ情報1102aや、当該ノードに対する差分データはリモートシステム110側に転送すべきかどうかを表わす本ノード有効フラグ1102bなどを含んでいる。

【0074】次に、上述のような構成の時系列キュー108へのエントリ手順の一例を、図13フローチャートを参照して説明する。ここでは、ある論理デバイス特定の論理トラックへの差分データに対するノードのキューへのエントリに対する例を挙げており、その更新データがマスタシステム100側で発生すると(ステップ1201)、そのデータの格納アドレス(キャッシュメモリ上アドレス、物理記憶媒体上(物理デバイス)アドレス)を求め(ステップ1202)、更新データをキャッシュメモリ104などに格納する(ステップ1203)。その後、本更新データに対する情報を組み合わせてノード情報を作成して(ステップ1204)、目的の時系列キュー108に接続する(ステップ1205)。更に、本論理トラックに対するフォーマット形式の差分データが既にあったか否かを調べ(ステップ1206)、あった場合には、以前に発生したフォーマットデータにあたる差分データはリモートシステム110側に転送する必要はないため、以前にエントリしたフォーマット形式の差分データに対するノードを検索し、そのノードに対する本ノード有効フラグ1102bをOFFにするか、あるいは時系列キュー108からはずすなどの手順を実施し(ステップ1207)、差分ビットマップへの反映は行わない。また、前記ステップ1206で、なかったと判定された場合には、差分ビットマップ107への反映を行う(ステップ1208)。

【0075】このような時系列キュー108を使用して(必要に応じて差分ビットマップを併用して)、回復コピー開始に伴い、その先頭位置にあるノードに対する差分データ、すなわち最も過去に発生した差分データから時系列に転送を開始することにより、整合性を維持すべき単位(論理デバイス単位など)での整合性を常に維持することが可能になる。

【0076】以上説明したように、本実施の形態のデータ多重化方法によれば、例えば「同期型」のデータ多重化を行う情報処理システム等において、なんらかの障害で「サスペンド状態」に陥った後の「回復コピー」実行中にマスタシステム100が重大災害によって運用不可な状態に陥ったとしても、リモートシステム110側のバックアップデータに対する整合性の保証を保持することができる為、リモートシステム110への運用移行を的確に行うことができる。

【0077】以上本発明者によってなされた発明を実施

の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【0078】たとえば、上述の実施の形態の説明では、一例としてマスタシステムおよびリモートシステムの二つのシステム間でデータを二重に保持させる例を説明したが、3以上のシステムにてデータ多重化を行う場合にも、本発明は適用できる。

【0079】

【発明の効果】本発明のデータ多重化方法によれば、複数のシステム間でのデータ多重化を維持するための差分データの回復複写における障害耐性を向上させることができる、という効果が得られる。

【0080】また、本発明のデータ多重化方法によれば、マスタシステムおよびリモートシステムが、ある整合性の維持を要求する情報単位において、特に「サスペンド状態」に陥る瞬間まで「ミラー状態」を維持していた場合の、「回復コピー」におけるリモートシステム側の多重化データの整合性を保証することができる、という効果が得られる。

【0081】また、本発明のデータ多重化方法によれば、マスタシステム側の障害等に際してリモートシステムに運用を切り替える上で運用可能な整合性の保証されたデータを提供することができる、という効果が得られる。

【図面の簡単な説明】

【図1】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムの構成の一例を示す概念図である。

【図2】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおけるRAID5のアーキテクチャの一例を示す概念図である。

【図3】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムでのバックアップコピーの手順の一例を図式化して例示した概念図である。

【図4】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムでのバックアップコピーの手順の一例を示すフローチャートである。

【図5】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおけるマスタシステムの作用の一例を示すフローチャートである。

【図6】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおけるリモートシステムの作用の一例を示すフローチャートである。

【図7】本発明の一実施の形態であるデータ多重化方法

が実施される情報処理システムにて使用される差分ビットマップの一例を示す概念図である。

【図8】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける回復コピー手順の一例を図式化して例示した概念図である。

【図9】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける回復コピー手順の一例を示すフローチャートである。

【図10】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける退避エリア管理ビットマップの一例を例示した概念図である。

【図11】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける退避データ情報キューの構造の一例を示す概念図である。

【図12】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける時系列キューの構造の一例を示す概念図である。

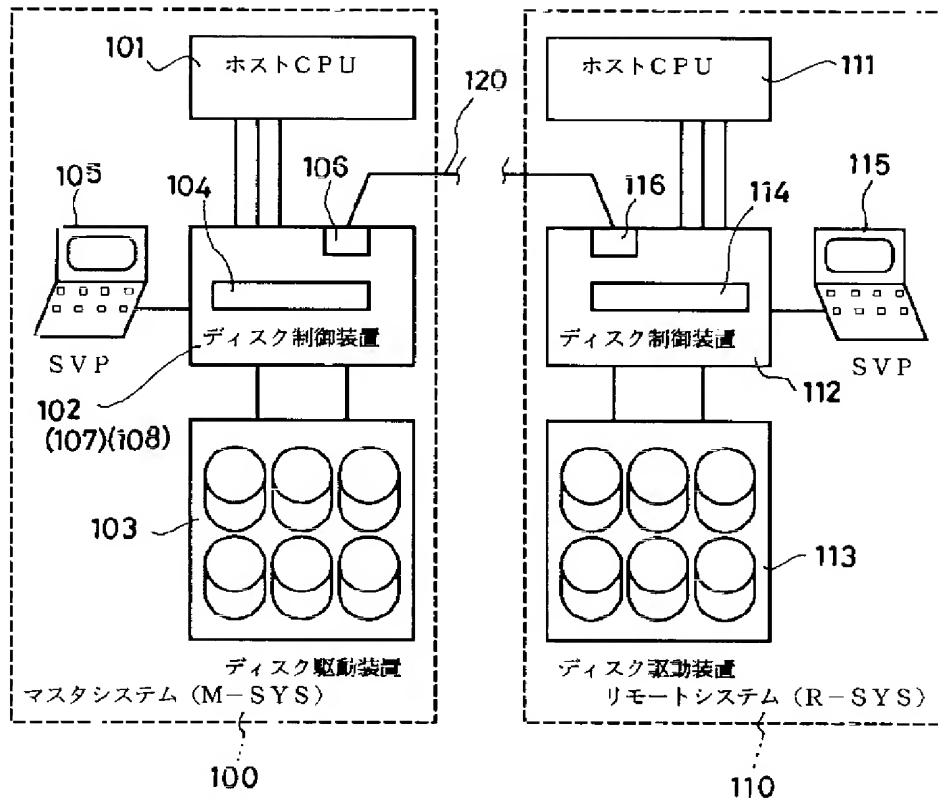
【図13】本発明の一実施の形態であるデータ多重化方法が実施される情報処理システムにおける時系列キューのエントリ手順の一例を示すフローチャートである。

【符号の説明】

100…マスタシステム（第1のシステム）、101…ホストCPU、102…ディスク制御装置、103…ディスク駆動装置（第1の記憶装置）、104…キャッシュメモリ、105…サービスプロセッサ、106…コネクタ、107…差分ビットマップ、108…時系列キュー、110…リモートシステム（第2のシステム）、111…ホストCPU、112…ディスク制御装置、113…ディスク駆動装置（第2の記憶装置）、114…キャッシュメモリ、115…サービスプロセッサ、116…コネクタ、117…システム管理用メモリ、118…退避エリア管理ビットマップ、119…退避データ情報キュー、120…データ転送路、200…小型ディスク、210…論理トラック、220…パリティトラック、230…スベアディスク、801…使用可否フラグ、802…ビットマップ、950…物理デバイス番号、951…物理トラック番号（セクタ番号）、952…ディスク番号、953…物理トラック番号（セクタ番号）、954…キャッシュメモリ上格納アドレス、1101…前エントリノードポインタ、1102…本ノード制御用情報、1102a…更新タイプ情報、1102b…本ノード有効フラグ、1103…論理トラック番号、1104…物理媒体上格納アドレス、1105…キャッシュメモリ上アドレス、1106…次エントリノードポインタ、T…回復コピー完了許容最大時間。

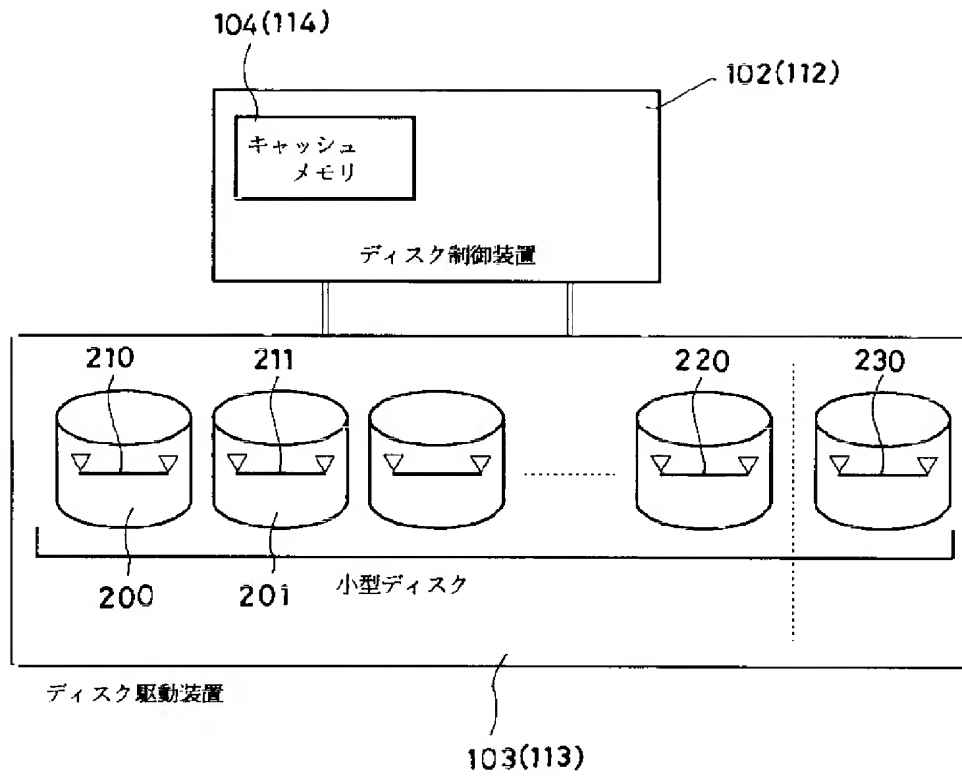
【図1】

図 1



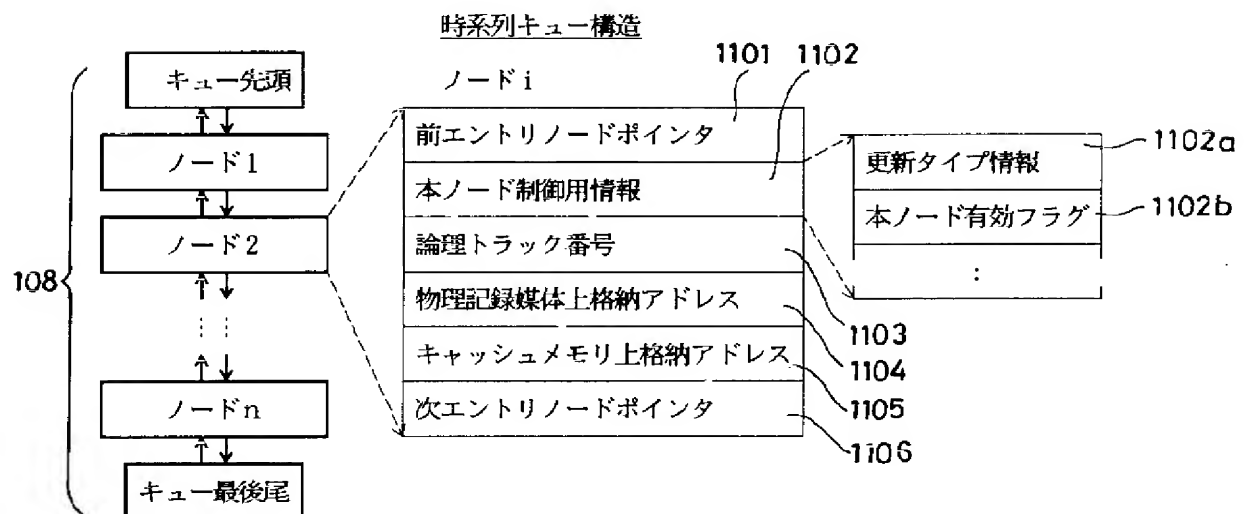
【図2】

図 2



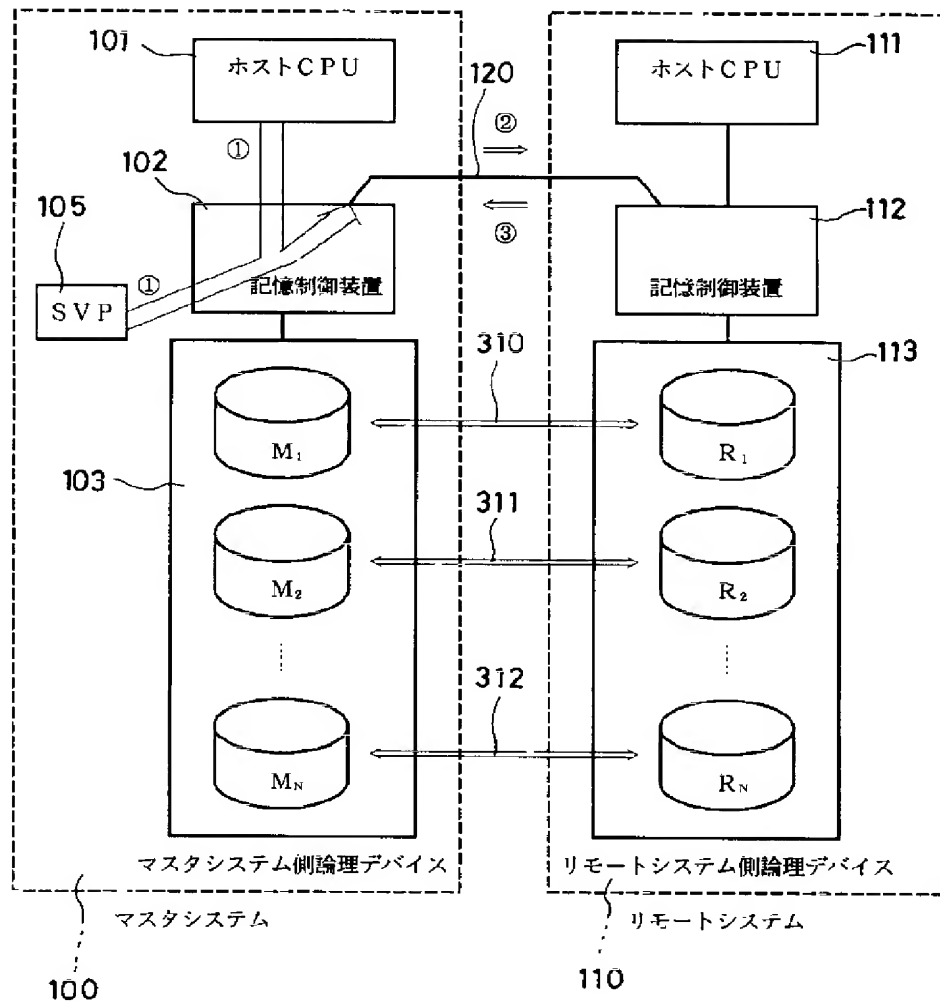
【図12】

図 1 2



【図3】

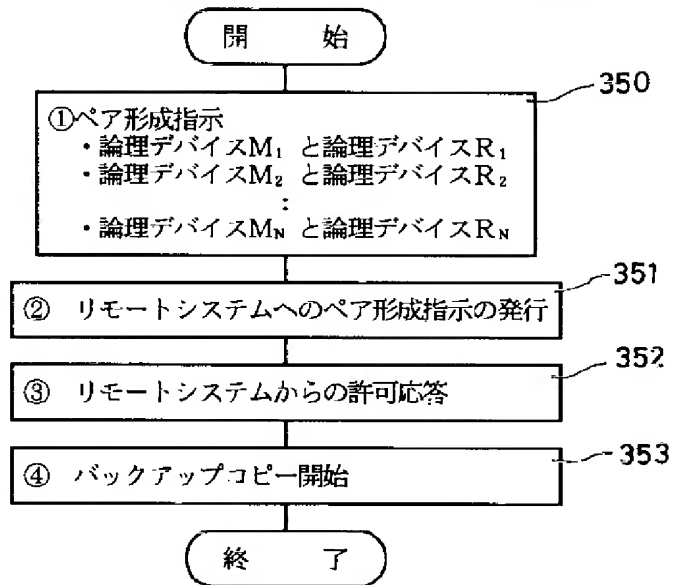
図3



【図4】

## 図4

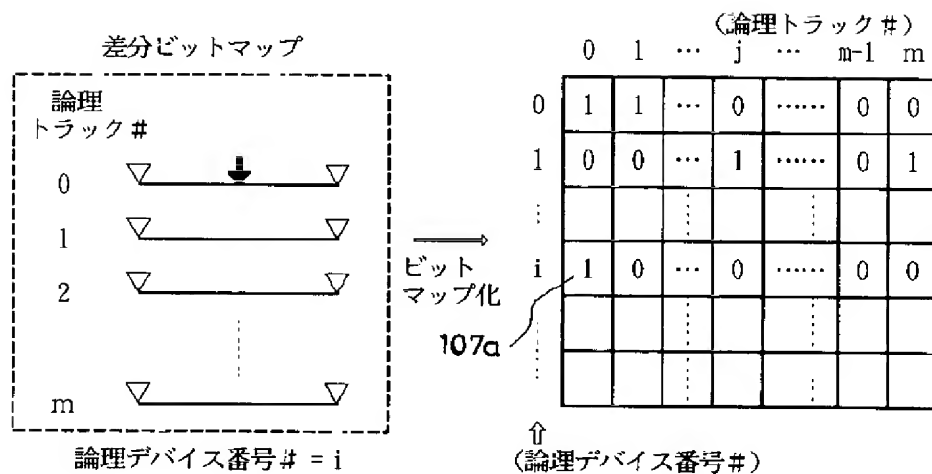
マスタ/リモートシステム間のバックアップコピー処理例



【図7】

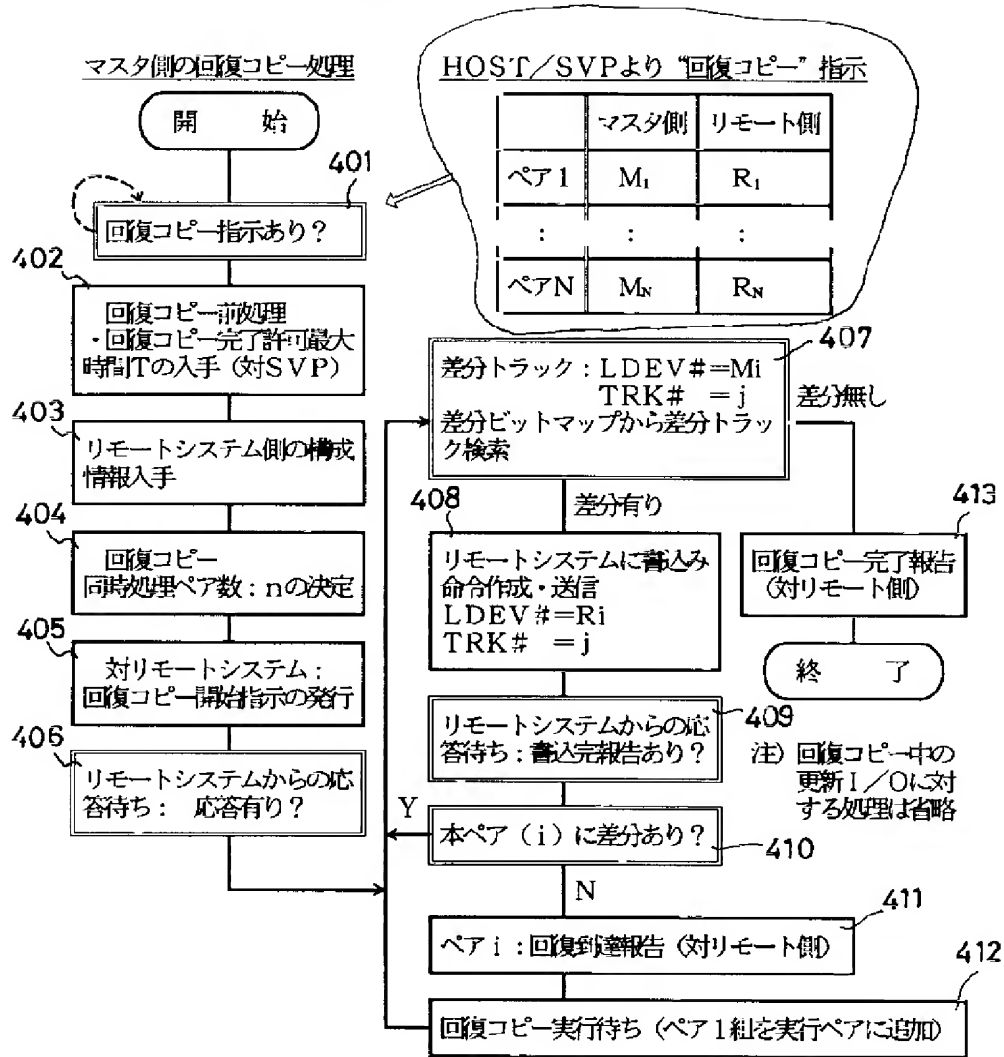
## 図7

107



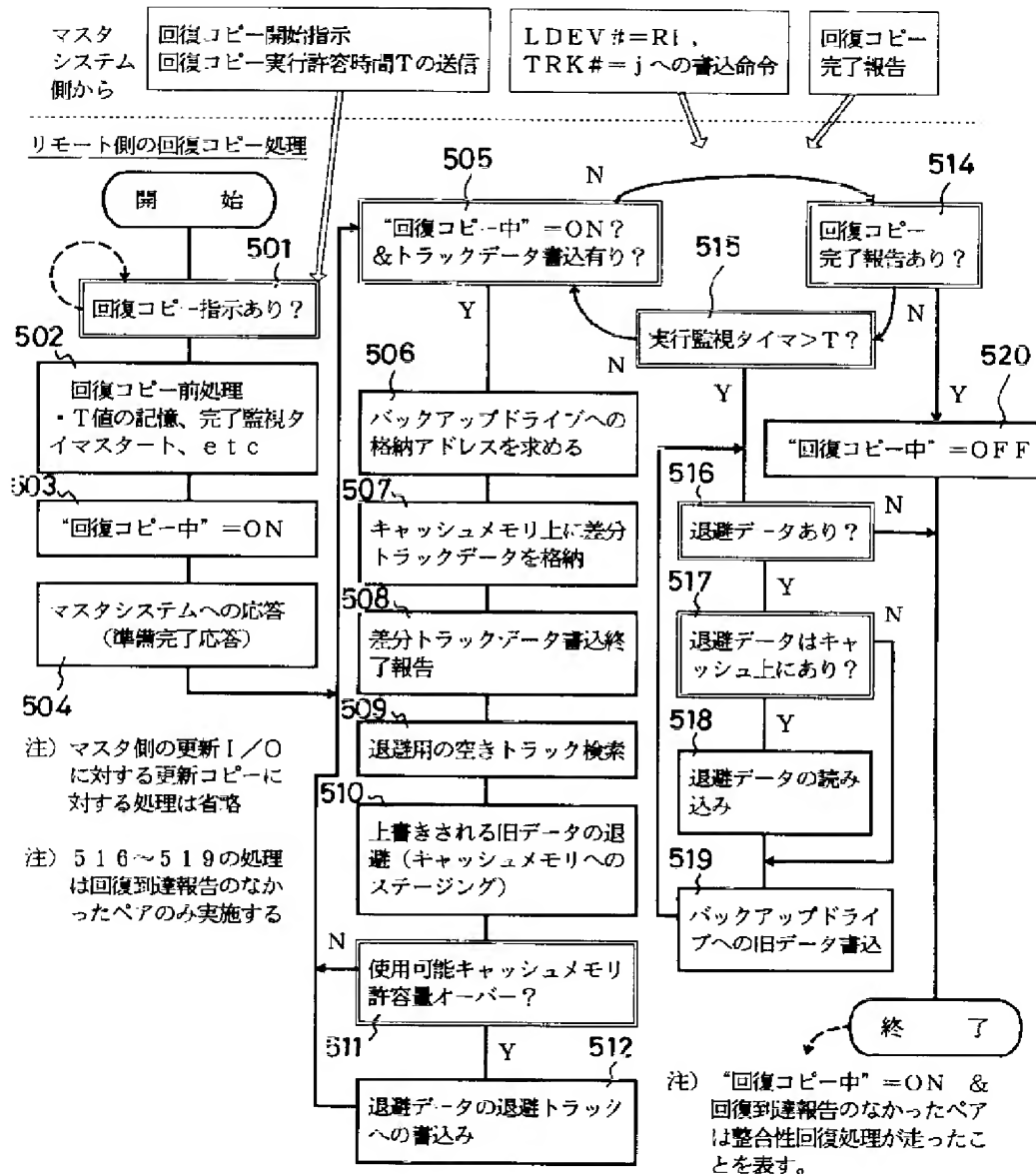
【図5】

図5



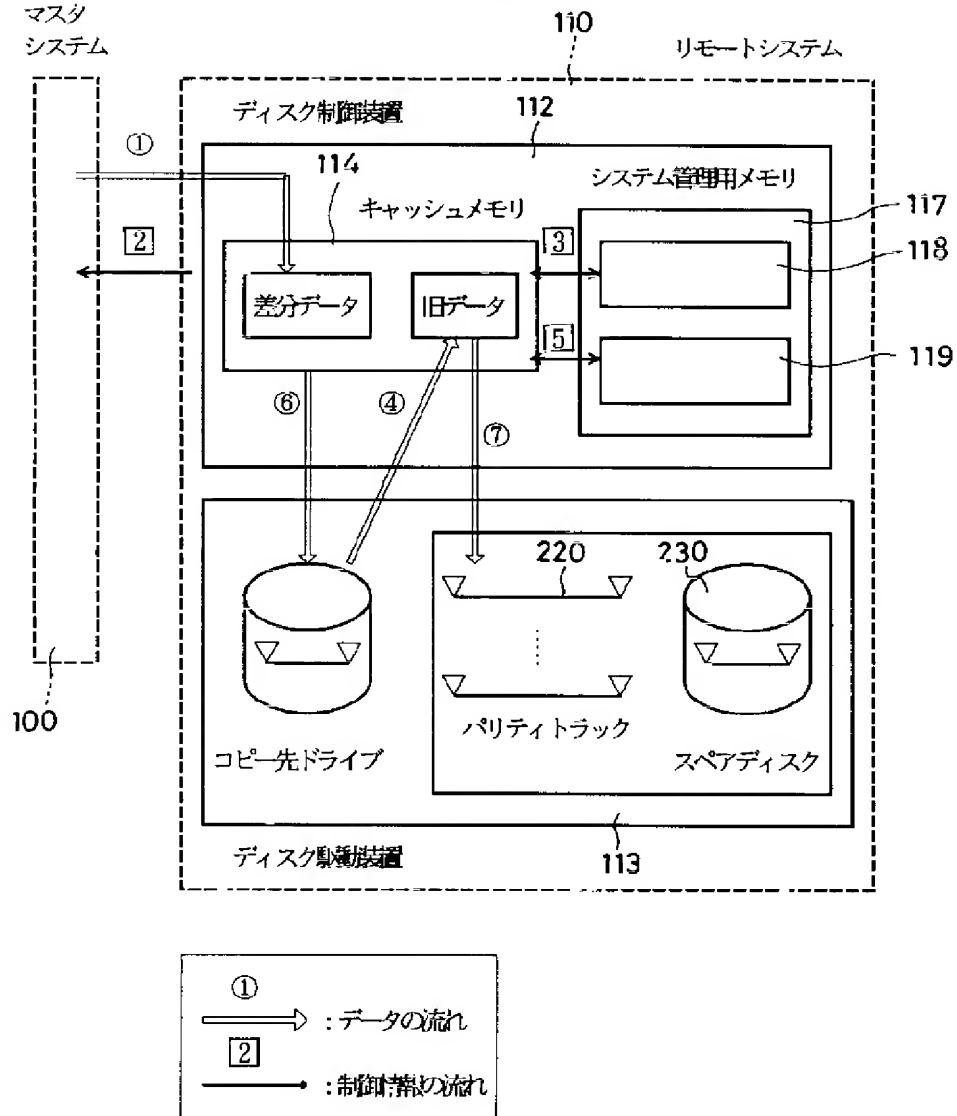
【図6】

図 6



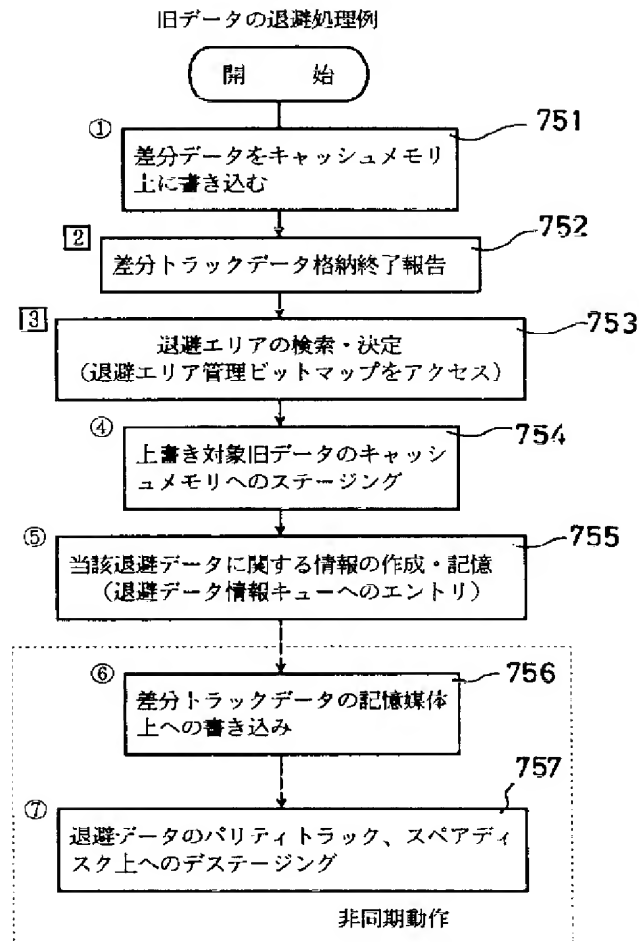
【図8】

図8



【図9】

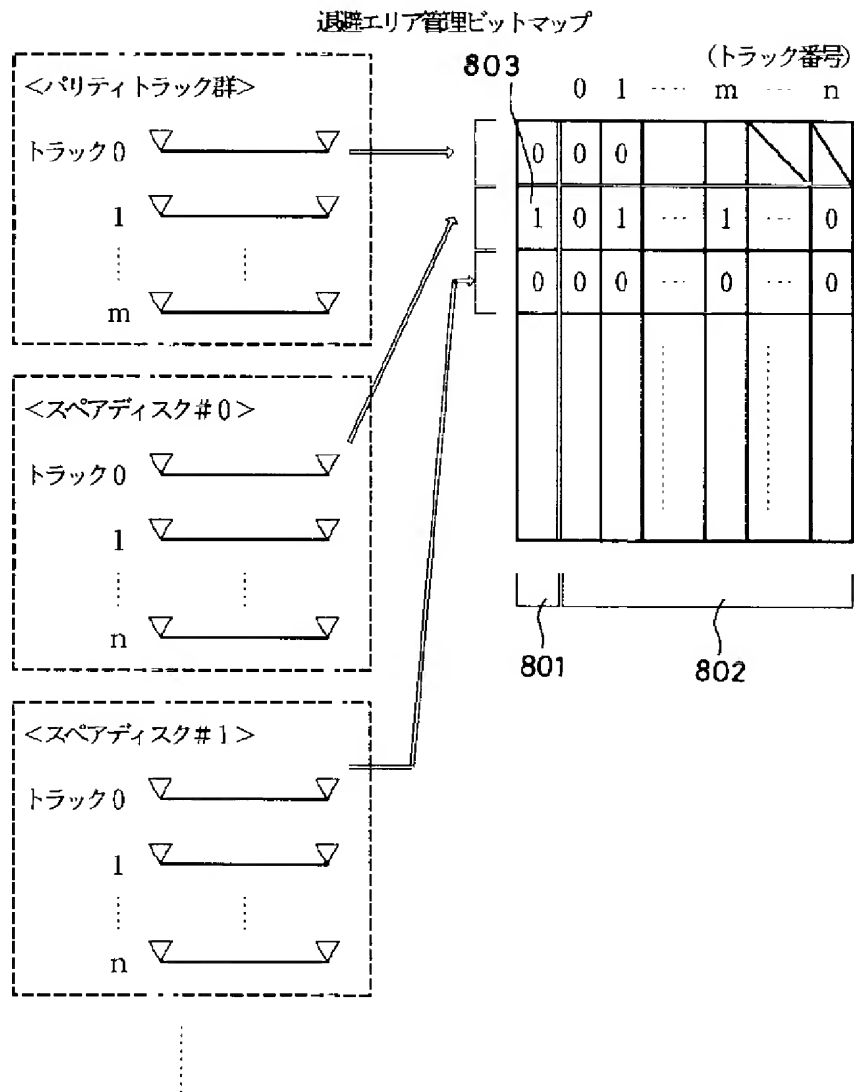
図 9



【図10】

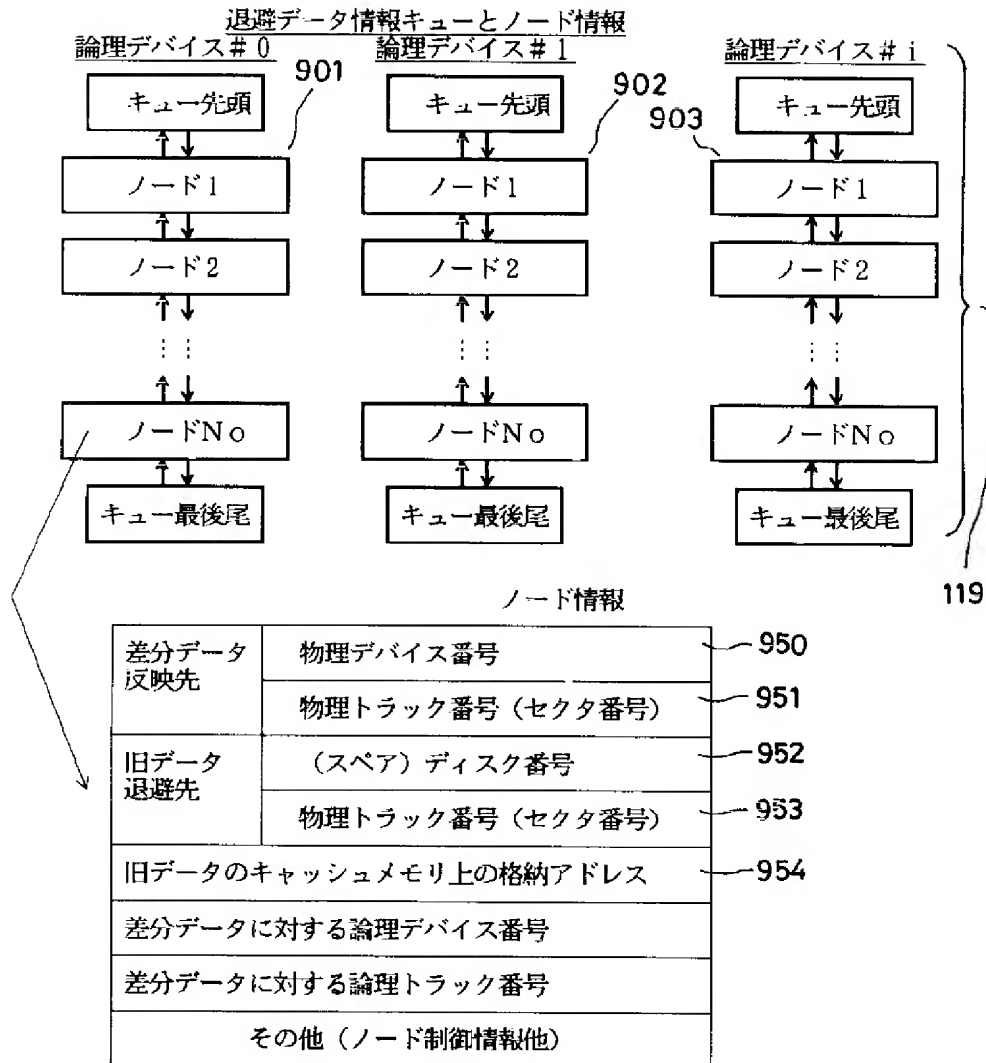
図10

118



【図11】

図 1 1



【図13】

## 図 1 3

